DOCUMENT RESUME

ED 093 957 TM 003 775

AUTHOR Gastright, Joseph F.
TITLE Some Empirical Evidence on the Stability of Discrepancy Measures Based on Observed and Predicted School Means on Achievement Tests.
PUB DATE Apr 74
NOTE 18p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, Illinois, April, 1974)

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS *Achievement Tests; *Correlation; *Elementary School Students; Longitudinal Studies; Models; Multiple Regression Analysis; *Predictive Ability (Testing); Predictor Variables; Reading Tests; Statistical Analysis
IDENTIFIERS Cincinnati Public Schools

ABSTRACT

An investigation was undertaken of the intercorrelations among residuals obtained by applying regression techniques under two conditions: (1) mean sixth-grade Metropolitan Achievement Test reading scores predicted by mean second-grade Stanford Achievement Test reading scores, and (2) mean sixth-grade reading scores predicted by mean second-grade reading scores and mobility variables. The predictions were performed on data for 66 elementary schools for three consecutive years. The addition of mobility variables significantly improved the prediction model. Intercorrelations among residuals within year and across models were uniformly high. (.77 to .94). Intercorrelations across years were not improved by the addition of mobility variables (.46 to .51). (Author)

SOME EMPIRICAL EVIDENCE ON THE STABILITY OF DISCREPANCY MEASURES BASED ON OBSERVED AND PREDICTED SCHOOL MEANS ON ACHIEVEMENT TESTS

by

Joseph F. Gastright
Department of Research & Development
Cincinnati Public Schools

Presented at annual meeting of the
American Educational Research Association
Chicago, Illinois

April, 1974

Some Empirical Evidence on the Stability of Discrepancy Measures Based on Observed and Predicted School Means on Achievement Tests.

Joseph F. Gastright
Cincinnati Public Schools

The conceptualization of school systems as input/output systems has led to the application of various mathematical models for predicting school outputs. Dyer (1966) proposed the use of multiple linear regression techniques for obtaining discrepancy measures based on observed and predicted school system achievement means. Dyer, Linn, and Patton (1969) presented empirical evidence on the comparability of discrepancy measures obtained using four different methods.

Method I utilized the regression of individual student output achievement on student input achievement using a student sample identical at two grade levels (matched longitudinal sample). Method II utilized the regression of mean school system output on mean school system input for the same matched-longitudinal sample. Method III utilized the regression of mean school system output on mean school system input for all students available at those points in time (unmatched-longitudinal sample). Method IV utilized regression of mean school system output on the concurrent school system mean of the earlier grade level (cross-sectional sample).

The "Student-Change Model" of an Educational System proposed by Dyer (1967) included among the independent variables not only input achievement but also measures of the hard to change conditions (e.g. students socioeconomic status, community wealth) and educational process variables (e.g. variety of teaching methods). Data on the surrounding condition and process variables were not available, however, to Dyer and his associates

in the data base of the New York study.

Dyer, et. al, using fifth grade achievement scores to predict eighth grade achievement scores (Iowa Test of Basic Skills), concluded that Methods I and II were essentially interchangeable, but not comparable to Methods III and IV. Operating under the assumption that the methods utilizing matched student samples were intrinsically superior to the others, they concluded that Methods III and IV did not produce residuals which were sufficiently comparable to those from Methods I and II to serve as reasonable substitutes for them. The stability of the residuals produced by Methods I and II was estimated by randomly dividing the student population into two groups and correlating the residuals produced by these two independent estimates. They reported a median correlation across the subtests of .70. This procedure randomizes the various factors that have contributed to differential educational programs and gives an estimate of the amount of error associated with pupils.

Forsyth (1973) provided some evidence on a different kind of stability, the consistency of residuals for consecutive classes in the same school. He randomly sampled 50 students from each of 320 schools in Iowa and utilizing Dyer's Method II, predicted mean school twelfth grade achievement scores using mean school ninth grade achievement scores (Iowa Test of Basic Skills). The multiple correlation coefficients reported by Forsyth are very consistent with those reported by Dyer. However, the intercorrelation between residuals for the consecutive years (median r=.28) were considerably lower than the random halves correlations reported by Dyer.

Acland (1972) reported somewhat higher intercorrelations among residuals (r=.4) for consecutive classes using unmatched student groups (Dyer's

Method III). Using longitudinal achievement test data from grades one to six (Metropolitan Achievement Test), Acland found that the intercorrelation between residuals for consecutive classes increased as the interval between the testing dates increased from one to three years. Acland interpreted the correlation between residuals as "a measure of the percentage of the variation in the residual scores that can be attributed to the stable characteristics of schools that raise performance level."

This study investigates the intercorrelations among residuals for three consecutive years using unmatched longitudinal student sample (Dyer's Method III). The results were obtained under two conditions: 1) mean school sixth grade reading achievement predicted by mean school second grade reading achievement, and 2) mean school sixth grade reading achievement predicted by mean school second grade reading achievement and school and community background variables collected concurrently with the second grade achievement testing. The multiple regressions and intercorrelation among residuals for three consecutive years were compared to determine whether 1) the addition of background variables significantly increased the predictive accuracy of the regression model, and 2) these factors increase the year to year stability of residuals obtained using the regression model.

Procedures

Longitudinal school mean data are available on 64 elementary schools in the Cincinnati Public School System. Second grade Stanford Achievement test means in reading comprehension, and sixth grade Metropolitan Achievement test means in paragraph meaning were available for all schools for three consecutive years (1967-71, 1968-72, 1969-73). The School Information System, a computerized data bank, routinely collects and reports a variety of data on schools in the system. The mobility and background variables

were utilized to develop the regression equations for this study.

Intercorrelations between mobility, background variables, and mean school reading achievement were used to limit the number of variables under consideration to fifteen. Stepwise multiple regression were then run to identify variables which contributed significantly to the prediction. To be included in the prediction equation a variable had to increase the squared multiple correlation significantly ($p < .05$). Because the final regression model was to be used to produce an interpretive report on school achievement for the benefit of principals and school system decision makers, certain non-statistical criteria were applied. The model building procedures specified by Draper and Smith (1966) were followed with the additional stipulation that the variables included in the final equation should have an educationally plausible relationship with reading achievement. Since the regression equations were to be used for predicting future achievement, not simply fitting the data, the variables entered had to show stability over time.

Analysis of the initial regression equations revealed two major problems. First, if all achievement test subtests were admitted to the independent variables, then no background or mobility variables contributed significantly to the prediction of the reading output scores. Furthermore, the pattern of significant achievement predictors over the three years was not stable. In subsequent analysis the achievement input was limited to the corresponding reading subtest. Second, when the achievement input was restricted, the mobility and background variables which contributed statistically to the prediction were not identical over the three years. The non-achievement variables, like the achievement subtests, tended to be highly inter-correlated. A number of different models were

were investigated (see Tables 1-4) for stability. The elimination of variables was based on both empirical and rational reasons.

The residuals obtained with various models were correlated within each year and across the three years. The regression equation selected for the final report (Three Variable Model 1971/72) was applied to the data from the other two years. Correlations were then run between these three sets of residuals to estimate the stability of a single model when applied to a different data base.

Results

The multiple correlation coefficients for the various models across all three years are between .85 and .93. The data reported for a single year (Tables 1-4) are representative of the results for the other two years. The variables included in the different models are the same across the three years. The regression coefficients are, of course, unique to the data within each year. When achievement test inputs are restricted, the mobility and background variables add significantly to the prediction of output reading achievement. The number of non-achievement variables contributing significantly to the prediction varied from one to four over the three years.

The correlation of residuals within year and across models were in the range .70 to .95 (Table 5). In general, the addition of variables lowers the correlation of the residuals with those based on achievement alone.

The correlations between residuals within the same model but across the three years (Table 6) ranged from .25 to .56. There is a systematic reduction in the correlation between residuals as the number of variables admitted to the model increase.

The three variable model for 1971/72 was applied to the input data for the other two years. The intercorrelations between residuals for the three years (Table 7) are fairly consistent and not surprisingly different from those obtained using the regression equations descriptively (Table 6).

PAGE 7

Discussion

The results of this study show that background and mobility variables can add significantly to the accuracy of multiple regression predictions of school output, when the number of achievement test inputs is restricted. The non-achievement variables available for this study were all moderately or highly intercorrelated, and their pattern of entry into the stepwise multiple regressions was not stable over the three years studied. The two non-achievement variables which showed the most stable behavior in our study were rather commonly reported school system descriptors (percent absence, percent transfers).

Since the residuals derived from the multiple regression analysis are to be interpreted as estimates of a school's performance, the question of selecting the appropriate model becomes a potentially explosive one. This raises an important question: to what extent are the residuals obtained within a given year affected by the selection of different groups of input variables? The correlations in Table 5 suggest that as the regression model is over-fitted the correlation of the residuals with those obtained from achievement input alone decreases. In absolute terms this can mean that the given school's actual mean achievement in reading can be described as significantly below expectation or moderately above expectation based solely on the choice of the regression model. Admittedly, these cases are not common but they do occur in our data.

The non-achievement variables used in our analyses were in many cases highly accurate (e.g. head counts) but not necessarily well behaved, in the sense of being normally distributed. This suggests that reliance on statistical probabilities alone in the selection of regression models would be unwise, particularly models based on a single year's data. The model verification procedures suggested by Draper and Smith proved useful in eliminating

potentially troublesome input variables. Two types of problems should be investigated carefully. First, some variables are subject to redefinition by policy decision or change in data collection. Second, some variables have spectacular "outliers" which can lead to spurious predicted outputs.

The correlations of residuals within model but across the three years provides no evidence that the addition of non-achievement variables increases the stability of residuals across years. On the contrary the results indicate that the residuals based on achievement input alone are more highly correlated than those obtained from the other models. The over-fitted models (those containing marginally significant input variables) are significantly less stable.

The consecutive year stabilities are much higher than those reported by Forsyth and marginally higher than those reported by Acland. The results obtained by applying a single regression equation to the data from the three years (Table 7) show highly consistent correlations among the residuals. Of course correlations among residuals do not indicate the absolute magnitude of the changes in either actual output, gain scores from second to sixth grade, or even differences between residuals.

The median differences in residuals for the same school for consecutive years were in the order of .5 years. These results and the variation in gain scores for the same schools over the same period indicate that there is substantial unexplained variation in school performance associated with a particular cohort. Acland found similar results in the New York Study. Since this variability cannot be associated with corresponding changes in the socioeconomic status of the parents in the school, it suggests that measurement of cohort parameters associated with a particular grade level, within a particular school (e.g. group values, leadership, etc.), might account for

a significant portion of school effectiveness. Since these variables are plausibly within the range of school influence such an investigation might be of considerable educational importance.

Estimates of stable school effects in the present study are comparable to those reported by Acland. While these estimates are smaller than some educators and sociologists would have predicted, this does not eliminate the possibility that changes in school program can increase the impact of education. As many have pointed out, the alternate hypothesis has still to be refuted; the hypothesis that on important parameters schools are very much alike. The development and use of a baseline model for educational output would potentially allow all schools to climb above their previous effectiveness. The constant reapplication of regression analysis to each year's data leads to the old "zero-sum" game where half of the schools are "below prediction".

The use of multiple regression methods for investigating school system performance or school performance is still a fairly untried, or at least unreported, activity. The results of existing empirical studies, differing as they do in grade level, achievement test, and methodology provide a fairly modest base for generalizing on the utility of the method. On the one hand, the method has proved particularly useful as a way of dealing with the large amounts of "messy data" collected about schools. Perhaps the most fruitful use of Dyer's proposal in the near future will be in decreasing the number of parameters considered important in raising school performance, as measured on standardized achievement tests. On the other hand, the insights gained from using predictive models in education cannot be converted into "control models" because some of the important variables are not subject to manipulation by educators or indeed the people in a free society. To

emphasize the tentative nature of the predictions made using our model, the resulting interpretive report was described as an achievement forecast, not an expectation.

Table 1. All Ten Variables as Predictors of Sixth Grade Reading Achievement (1971/72)

Variable	Multiple R	Correlation with Criterion
Percent Transfers Out	.66	.66
Pupil/Parent Factor	.88*	.80
Second Grade Reading	.91*	.79
Percent on Welfare	.92	-.77
Percent Absence	.92*	-.61
Percent Transfers In	.93*	-.78
Pupil/Teacher Ratio	.93	.70
Number of Registered Voters	.93	.76
Number of Free Lunches	.93	-.54
Percent Voting	.93	-.60

*$p < .05$

Table 2. The Five Predictor Model of Sixth Grade Reading Achievement (1971/72)

Variable	Multiple	Correlation with Criterion
Second Grade Reading	.79*	.79
Percent on Welfare	.88*	-.77
Percent Transfers In	.90*	-.78
Pupil/Parent Factor	.90	.80
Percent Absence	.90	-.61

*$p < .05$

Table 3. The Four Best Predictor Model of Sixth Grade Reading Achievement (1971/72)

Variable	Multiple R	Correlation with Criterion
Second Grade Reading	.79	.79
Percent on Welfare	.88	-.77
Percent Absence	.89	-.61
Pupil/Parent Factor	.89	.80

*p < .05

Table 4. The Three Predictor Model of Sixth Grade Reading Achievement.

Variable	Multiple R	Correlation with Criterion
Second Grade Reading	.79*	.79
Percent Transfers In	.84*	-.78
Percent Absence	.86*	-.61

*p < .05

Table 5. Zero Order Correlations Between Residuals Within Year, Across Prediction Model (N=64)

Year 1970/71

	One	Three	Four	Five
Achievement	--			
Three Variable	.82	--		
Four Variable	.83	.95	--	
Five Variable	.77	.93	.94	--

Year 1971/72

Achievement	--			
Three Variable	.77	--		
Four Variable	.82	.85	--	
Five Variable	.73	.77	.90	--

Year 1972/73

Achievement	--			
Three Variable	.91	--		
Four Variable	.76	.84	--	
Five Variable	.70	.77	.92	--

Table 6. Zero Order Correlations Between Residuals Within Prediction Model, Across Years (N=64)

Achievement Alone

	70/71	71/72	72/73
70/71	--		
71/72	.56	--	
72/73	.46	.54	--

Three Variable Prediction Model

	70/71	71/72	72/73
70/71	--		
71/72	.51	--	
72/73	.47	.46	--

Four Variable Model

	70/71	71/72	72/73
70/71	--		
71/72	.51	--	
72/73	.42	.37	--

Five Variable Model

	70/71	71/72	72/73
70/71	--		
71/72	.42	--	
72/73	.39	.25	--

Table 7. Zero Order Correlations Between Residuals Obtained by Applying the Best Three Variable Model (1971/72) To Data for all Three Years. (N=66)

	70/71	71/72	72/73
70/71	--		
71/72	.46	--	
72/73	.41	.44	--

REFERENCES

1. Acland, H., The Impact of Schooling in the Elementary Grades. The Center for Educational Policy Research, Harvard Graduate School of Education, 1972.

2. Brophy, J.E., Stability in Teacher Effectiveness. Research and Development Report Series 77. The Research and Development Center for Teacher Education, University of Texas at Austin, 1972.

3. Forsyth, R.A., Some Empirical results related to the stability of performance indicators in Dyer's student change model of an educational system, Journal of Educational Measurement, 1973, 10, 7-12.

4. Draper, N.R., and Smith, H. Applied Regression Analysis, John Wiley & Sons: New York, 1966.

5. Dyer, H.S., The Pennsylvania Plan, Science Foundation, 1966, 50, 242-248.

6. Dyer, H.S., Linn, R.L., and Patton, M.J., Feasibility study of educational performance indicators: final report to New York State Education Department., Princeton: Educational Testing Service, 1967.

7. Dyer, H.S., Linn, R.L., Patton, M.J., A Comparison of four methods of obtaining discrepancy measures based on observed and predicted school system means on achievement tests, American Educational Research Journal, 1969, 6, 591-606.

8. Soar, R.S., Accountability: problems and possibilities in accountability as the measurement of pupils. A paper presented at the American Educational Researchers Association Meeting, 1973.